Decision Tree with Continuous Variables

Prepared by Dr. Suneesh Jacob

Dataset

The dataset contains three attributes: **Age**, **Income**, and the **Purchase** decision (target variable).

S.No.	Age	Income	Purchase (Target)
1	22.5	35.2	0
2	25.0	40.5	0
3	30.5	50.0	1
4	35.7	60.3	1
5	40.0	70.1	0
6	45.2	80.0	1

Let us choose all possible mid-lines of the points to split.
If we take the column 'Age', we can have these possible splits:

S.No.	Split Point	Left Split	Right Split
1	$\frac{22.5+25.0}{2} = 23.75$	Age <= 23.75	Age > 23.75
2	$\frac{25.0+30.5}{2} = 27.75$	Age <= 27.75	Age > 27.75
3	$\frac{30.5+35.7}{2} = 31.1$	Age <= 31.1	Age > 31.1
4	$\frac{35.7+40.0}{2} = 37.85$	Age <= 37.85	Age > 37.85
5	$\frac{40.0+45.2}{2} = 42.6$	Age <= 42.6	Age > 42.6

If we take the column 'Income', we can have these possible splits:

S.No.	Split Point	Left Split	Right Split

1	37.85	Income <= 37.85	Income > 37.85
2	45.25	Income <= 45.25	Income > 45.25
3	55.15	Income <= 55.15	Income > 55.15
4	65.2	Income <= 65.2	Income > 65.2
5	75.05	Income <= 75.05	Income > 75.05

Entropy of the Dataset

The total entropy of the dataset (before split) is:

$$E_{\text{Parent}} = -p_{\text{No}} \log_2 (p_{\text{No}}) - p_{\text{Yes}} \log_2 (p_{\text{Yes}}) = -\left(\frac{3}{6} \log_2 \frac{3}{6}\right) - \left(\frac{3}{6} \log_2 \frac{3}{6}\right) = 1$$

Information Gain for Age and Income Splits

The following table summarizes the entropy, weighted entropy, and information gain for all possible splits on **Age** and **Income**.

Feature	Split Point	Entropy of Left Split	Entropy of Right Split	Weighted Entropy Calculation	Information Gain
Age	23.75	$-\left(\frac{1}{1} \log_2 \frac{1}{1}\right) - \left(\frac{0}{1} \log_2 \frac{0}{1}\right) = 0$	$-\left(\frac{2}{5} \log_2 \frac{2}{5}\right) - \left(\frac{3}{5} \log_2 \frac{3}{5}\right) \approx 0.97$	$\frac{1}{6} \times 0 + \frac{5}{6} \times 0.97 \approx 0.81$	1-0.81 = 0.19
Age	27.75	$-\left(\frac{2}{2} \log_2 \frac{2}{2}\right) - \left(\frac{0}{2} \log_2 \frac{0}{2}\right) = 0$	$-\left(\frac{1}{4} \log_2 \frac{1}{4}\right) - \left(\frac{3}{4} \log_2 \frac{3}{4}\right) \approx 0.81$	$\frac{2}{6} \times 0 + \frac{4}{6} \times 0.81 \approx 0.54$	1-0.54 = 0.46
Age	33.10	$-\left(\frac{2}{3} \log_2 \frac{2}{3}\right) - \left(\frac{1}{3} \log_2 \frac{1}{3}\right) \approx 0.92$	$-\left(\frac{1}{3} \log_2 \frac{1}{3}\right) - \left(\frac{2}{3} \log_2 \frac{2}{3}\right) \approx 0.92$	$\frac{3}{6} \times 0.92 + \frac{3}{6} \times 0.92 = 0.92$	1-0.92 = 0.08
Age	37.85	$-\left(\frac{2}{4} \log_2 \frac{2}{4}\right) - \left(\frac{2}{4} \log_2 \frac{2}{4}\right) = 1$	$-\left(\frac{1}{2} \log_2 \frac{1}{2}\right) - \left(\frac{1}{2} \log_2 \frac{1}{2}\right) = 1$	$\frac{4}{6} \times 1 + \frac{2}{6} \times 1 = 1$	1-1 = 0.00
Age	42.60	$-\left(\frac{3}{5} \log_2 \frac{3}{5}\right) - \left(\frac{2}{5} \log_2 \frac{2}{5}\right) \approx 0.97$	$-\left(\frac{0}{1} \log_2 \frac{0}{1}\right) - \left(\frac{1}{1} \log_2 \frac{1}{1}\right) = 0$	$\frac{5}{6} \times 0.97 + \frac{1}{6} \times 0 = 0.81$	1-0.81=0.19
Income	37.85	$-\left(\frac{1}{1} \log_2 \frac{1}{1}\right) - \left(\frac{0}{1} \log_2 \frac{0}{1}\right) = 0$	$-\left(\frac{2}{5} \log_2 \frac{2}{5}\right) - \left(\frac{3}{5} \log_2 \frac{3}{5}\right) \approx 0.97$	$\frac{1}{6} \times 0 + \frac{5}{6} \times 0.97 \approx 0.81$	1-0.81=0.19
Income	45.25	$-\left(\frac{2}{2} \log_2 \frac{2}{2}\right) - \left(\frac{0}{2} \log_2 \frac{0}{2}\right) \approx 0$	$-\left(\frac{1}{4} \log_2 \frac{1}{4}\right) - \left(\frac{3}{4} \log_2 \frac{3}{4}\right) \approx 0.81$	$\frac{2}{6} \times 0 + \frac{4}{6} \times 0.81 = 0.54$	1-0.54=0.46
Income	55.15	$-\left(\frac{2}{3} \log_2 \frac{2}{3}\right) - \left(\frac{1}{3} \log_2 \frac{1}{3}\right) = 0.92$	$-\left(\frac{1}{3} \log_2 \frac{1}{3}\right) - \left(\frac{2}{3} \log_2 \frac{2}{3}\right) = 0.92$	$\frac{3}{6} \times 0.92 + \frac{3}{6} \times 0.92 = 0.92$	1-0.92=0.08
Income	65.20	$-\left(\frac{2}{4} \log_2 \frac{2}{4}\right) - \left(\frac{2}{4} \log_2 \frac{2}{4}\right) \approx 1$	$-\left(\frac{1}{2} \log_2 \frac{1}{2}\right) - \left(\frac{1}{2} \log_2 \frac{1}{2}\right) = 1$	$\frac{4}{6} \times 1 + \frac{2}{6} \times 1 = 1$	1-1=0.00
Income	75.05	$-\left(\frac{3}{5} \log_2 \frac{3}{5}\right) - \left(\frac{2}{5} \log_2 \frac{2}{5}\right) = 0.97$	$-\left(\frac{0}{1} \log_2 \frac{0}{1}\right) - \left(\frac{1}{1} \log_2 \frac{1}{1}\right) = 0$	$\frac{5}{6} \times 0.97 + \frac{1}{6} \times 0 = 0.81$	1-0.81=0.19

Summary of Information Gain

The best split point based on information gain for each feature is as follows:

Best split for Age: Age = 27.75 with an information gain of 0.46.
Best split for Income: Income = 45.25 with an information gain of 0.46.
Best split for the entire possible splits: Either Age = 27.75 or Income = 45.25, since both have the same Information Gain.

Since there is a tie, let us go ahead to break the tie by arbitrarily choosing Age = 27.75 as the splitting criterion.
The resultant branched datasets are shown in the tables below.

S.No.	Age	Income	Purchase (Target)
1	22.5	35.2	0
2	25.0	40.5	0

Branch 1

S.No.	Age	Income	Purchase (Target)
3	30.5	50.0	1
4	35.7	60.3	1
5	40.0	70.1	0
6	45.2	80.0	1

Branch 2

The current tree is shown in Figure 1.
In Branch 1, all the output labels are of the same class. Hence, this would amount to a leaf node. In Branch 2, not all the labels are of the same class. Hence, further splitting is required for this node.

Information Gain for Age and Income Splits

The total entropy of the Branch 2 dataset (before split) is:

$$E_{\text{Parent}} = -\left(\frac{1}{4}\log_2\frac{1}{4}\right) - \left(\frac{3}{4}\log_2\frac{3}{4}\right) \approx 0.81$$

The following table summarizes the entropy, weighted entropy, and information gain for all possible splits on **Age** and **Income**.

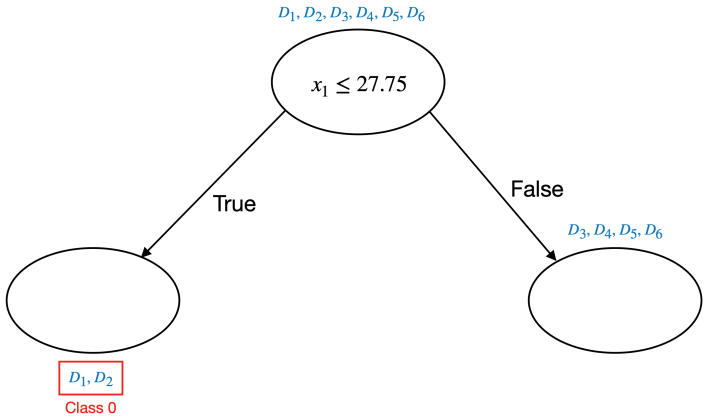

Figure 1: Tree at first iteration

Feature	Split Point	Entropy of Left Split	Entropy of Right Split	Weighted Entropy Calculation	Information Gain
Age	33.10	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{1}{1}\log_2\frac{1}{1}\right)=0$	$-\left(\frac{1}{3}\log_2\frac{1}{3}\right)-\left(\frac{2}{3}\log_2\frac{2}{3}\right)\approx 0.92$	$\frac{1}{4}\times 0+\frac{3}{4}\times 0.92\approx 0.69$	0.81-0.69 = 0.12
Age	37.85	$-\left(\frac{0}{2}\log_2\frac{0}{2}\right)-\left(\frac{2}{2}\log_2\frac{2}{2}\right)=0$	$-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)=1$	$\frac{2}{4}\times 0+\frac{2}{4}\times 1=0.5$	0.81-0.5 = 0.31
Age	42.60	$-\left(\frac{1}{3}\log_2\frac{1}{3}\right)-\left(\frac{2}{3}\log_2\frac{2}{3}\right)\approx 0.92$	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{1}{1}\log_2\frac{1}{1}\right)=0$	$\frac{3}{4}\times 0.92+\frac{1}{4}\times 0=0.69$	0.81-0.69=0.12
Income	55.15	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{1}{1}\log_2\frac{1}{1}\right)=0$	$-\left(\frac{1}{3}\log_2\frac{1}{3}\right)-\left(\frac{2}{3}\log_2\frac{2}{3}\right)\approx 0.92$	$\frac{1}{4}\times 0+\frac{3}{4}\times 0.92\approx 0.69$	0.81-0.69 = 0.12
Income	65.20	$-\left(\frac{0}{2}\log_2\frac{0}{2}\right)-\left(\frac{2}{2}\log_2\frac{2}{2}\right)=0$	$-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)=1$	$\frac{2}{4}\times 0+\frac{2}{4}\times 1=0.5$	0.81-0.5 = 0.31
Income	75.05	$-\left(\frac{1}{3}\log_2\frac{1}{3}\right)-\left(\frac{2}{3}\log_2\frac{2}{3}\right)\approx 0.92$	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right)-\left(\frac{1}{1}\log_2\frac{1}{1}\right)=0$	$\frac{3}{4}\times 0.92+\frac{1}{4}\times 0=0.69$	0.81-0.69=0.12

Summary of Information Gain

The best split point based on information gain for each feature is as follows:

Best split for Age: Age = 37.85 with an information gain of 0.31.

Best split for Income: Income = 65.20 with an information gain of 0.31.

Best split for the entire possible splits: Either Age = 37.85 or Income = 65.20, since both have the same Information Gain.

Since there is a tie, let us go ahead to break the tie by arbitrarily choosing Age = 37.85 as the splitting criterion.
The resultant branched datasets are shown in the tables below.

S.No.	Age	Income	Purchase (Target)
3	30.5	50.0	1
4	35.7	60.3	1

Branch A

S.No.	Age	Income	Purchase (Target)
5	40.0	70.1	0
6	45.2	80.0	1

Branch B

The current tree is shown in Figure 2.

In Branch A, all the output labels are of the same class. Hence, this would amount to a leaf node. In Branch B, not all the labels are of the same class. Hence, further splitting is required for this node.

Information Gain for Age and Income Splits

The total entropy of the Branch B dataset (before split) is:

$$E_{\text{Parent}} = -\left(\frac{1}{2}\log_2\frac{1}{2}\right) - \left(\frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

The following table summarizes the entropy, weighted entropy, and information gain for all possible splits on **Age** and **Income**.

Feature	Split Point	Entropy of Left Split	Entropy of Right Split	Weighted Entropy Calculation	Information Gain
Age	42.60	$-\left(\frac{1}{1}\log_2\frac{1}{1}\right) - \left(\frac{0}{1}\log_2\frac{0}{1}\right) = 0$	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right) - \left(\frac{1}{1}\log_2\frac{1}{1}\right) = 0$	$\frac{1}{2}\times 0 + \frac{1}{2}\times 0 = 0$	1-0=1
Income	75.05	$-\left(\frac{1}{1}\log_2\frac{1}{1}\right) - \left(\frac{0}{1}\log_2\frac{0}{1}\right) = 0$	$-\left(\frac{0}{1}\log_2\frac{0}{1}\right) - \left(\frac{1}{1}\log_2\frac{1}{1}\right) = 0$	$\frac{1}{2}\times 0 + \frac{1}{2}\times 0 = 0$	1-0=1

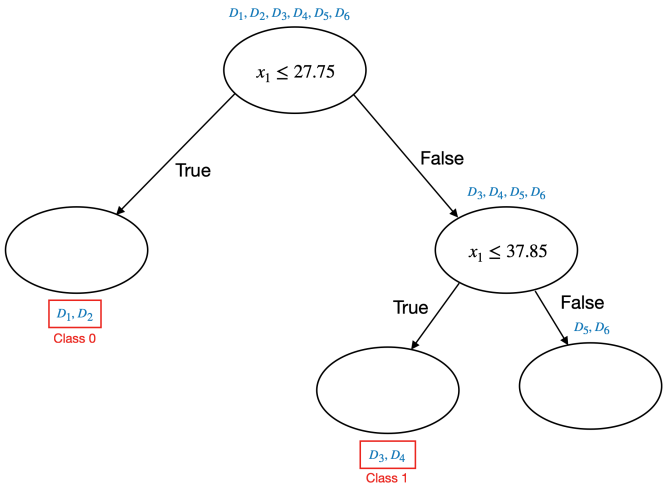

Figure 2: Tree at second iteration

Summary of Information Gain

The best split point based on information gain for each feature is as follows:

Best split for Age: Age = 42.60 with an information gain of 1.

Best split for Income: Income = 75.05 with an information gain of 1.

Best split for the entire possible splits: Either Age = 42.60 or Income = 75.05, since both have the same Information Gain.

Since there is a tie, let us go ahead to break the tie by arbitrarily choosing Age = 42.60 as the splitting criterion.

The resultant branched datasets are shown in the tables below.

S.No.	Age	Income	Purchase (Target)
5	40.0	70.1	0

Branch P

S.No.	Age	Income	Purchase (Target)
6	45.2	80.0	1

Branch Q

The current tree is shown in Figure 3.

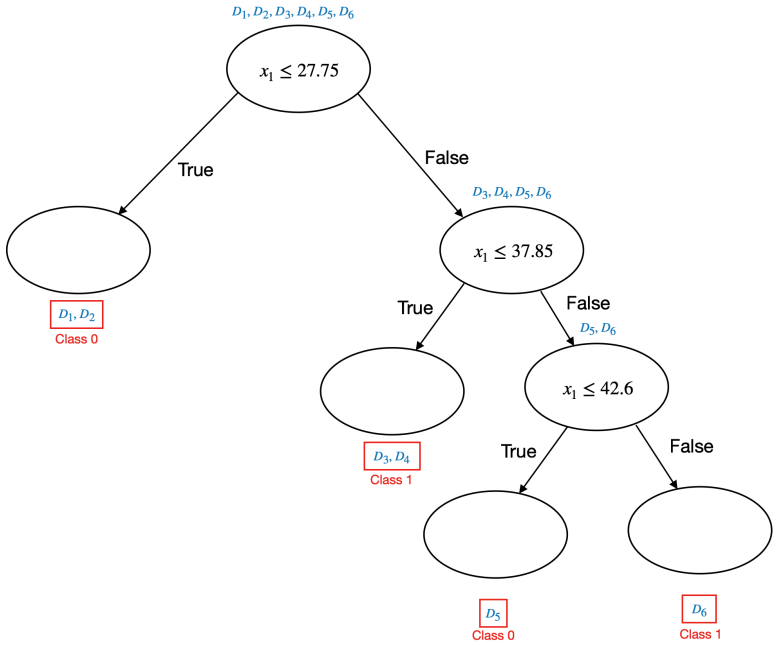

Figure 3: Tree at third iteration

For both branches P and Q, all the output labels are of the same class. Hence, both would amount to a leaf nodes and no further splitting is required. The final decision tree is as shown in Figure 3.